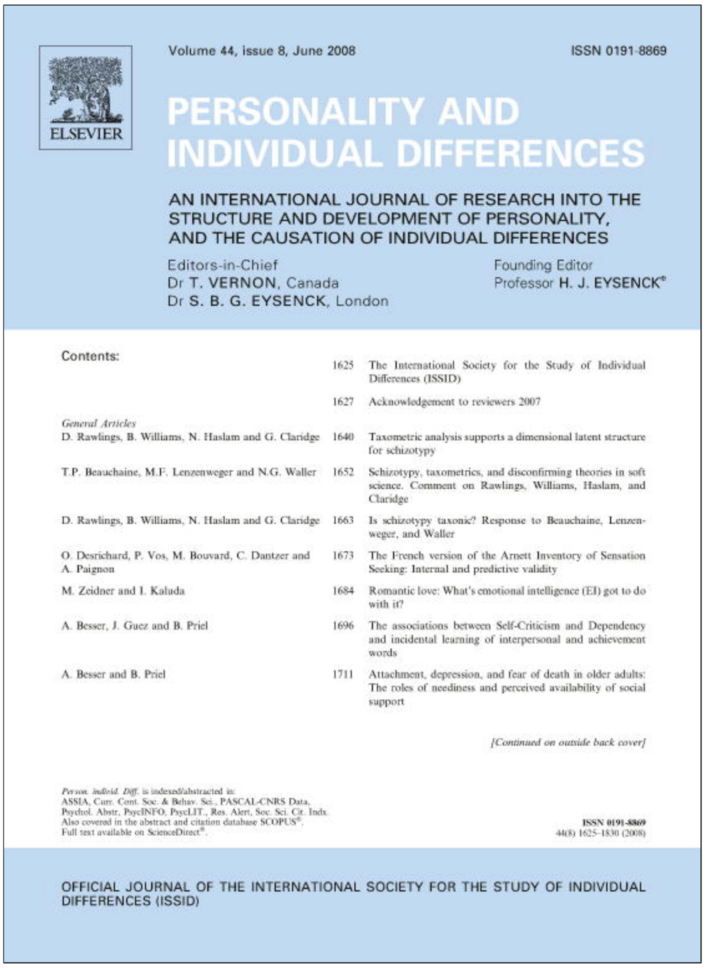

Volume 44, issue 8, June 2008

ISSN 0191-8869

ELSEVIER

PERSONALITY AND INDIVIDUAL DIFFERENCES

AN INTERNATIONAL JOURNAL OF RESEARCH INTO THE STRUCTURE AND DEVELOPMENT OF PERSONALITY, AND THE CAUSATION OF INDIVIDUAL DIFFERENCES

Contents:

General Articles

[Continued on outside back cover]

Person. individ. Diff. is indexed/abstracted in:
ASSIA, Curr. Cont. Soc. & Behav. Sci., PASCAL-CNRS Data, Psychol. Abstr, PsycINFO, PsycLIT., Res. Alert, Soc. Sci. Cit. Indx.
Also covered in the abstract and citation database SCOPUS®.
Full text available on ScienceDirect®.

ISSN 0191-8869
44(8) 1625–1830 (2008)

OFFICIAL JOURNAL OF THE INTERNATIONAL SOCIETY FOR THE STUDY OF INDIVIDUAL DIFFERENCES (ISSID)

Available online at www.sciencedirect.com

ScienceDirect

Personality and Individual Differences 44 (2008) 1819–1823

PERSONALITY AND INDIVIDUAL DIFFERENCES

www.elsevier.com/locate/paid

Short Communication

Not all effect sizes are the same: Comments on Holden (2008)

Robert E. McGrath *

School of Psychology, T-WH1-01, Fairleigh Dickinson University, Teaneck, NJ 07666, United States

Received 30 November 2007; received in revised form 15 January 2008; accepted 29 January 2008
Available online 17 March 2008

Abstract

Despite the common belief that response bias is a significant moderator of psychological tests in field settings, these biases have been notoriously difficult to identify. Holden (2008) has recently presented evidence suggesting this paradox may at least in part be explained by problems inherent to the use of moderated regression with self-report indicators of response bias. His article offers an innovative proposal for understanding a central issue in applied test use. However, the conclusions drawn about both moderated regression and the general validity of response bias indicators are open to alternative explanations. It would be premature to assume these factors are important contributors to the ephemeral character of response bias effects.

Keywords: Faking; Response bias; Moderated regression; Effect sizes; Self-report

1. Introduction

After more than 60 years of analytic discussion (Cronbach, 1946), the role of faking and response bias in self-report measurement continues to vex psychologists interested in assessment. Despite widespread concern about the prevalence of invalid responding in applied settings (e.g., Gouvier, Lees-Haley, & Hammer, 2003; Mittenberg, Patton, Canyock, & Condit, 2002), a number of authors have concluded the importance of response bias has been greatly exaggerated (e.g., Hogan, Barrett, & Hogan, 2007; Ones & Viswesvaran, 1998; Piedmont, McCrae, Riemann, &

* Tel.: +1 201 692 2445; fax: +1 201 692 2304.
E-mail address: mcgrath@fdu.edu

0191-8869/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2008.01.015

Angleitner, 2000; Rorer, 1965). Holden (2008) brings a fresh approach to this issue, suggesting that the failure to find response bias effects may be a problem with commonly used statistical methods and bias indicators. Because it is a potentially important contribution to a very important literature, Holden's case deserves close consideration.

The first study described by Holden (2008) focused on psychopathic tendencies, the second on the five factor model (Costa & McCrae, 1991). In both studies, undergraduates completed one or more predictor scales and response bias indicators under either a standard, fake-good, or fake-bad instructional set. Three statistical models were used to detect the effect due to response bias. The first computed the difference in the proportion of criterion variance predicted by the predictor scale (the difference in the squared correlation) under standard instructions versus each faking instructional set. This will be referred to here as Model 1. These differences in the proportion of variance predicted varied between 10.96% and 16.70%, with a mean value weighted by sample size of 15.09%.

The second approach (Model 2) used moderated regression analyses in which the predictors were the predictor scale, a dichotomous variable representing membership in either the standard instruction group or one of the faking instruction groups, and the product term. The proportion of variance predicted by this last term, represented by the part correlation, varied between 0.29% and 11.63% with a mean of 5.17%. For Model 3, dimensional score on a response bias indicator was substituted for the dichotomous indicator in the moderated regression. The proportion of variance predicted declined even further, varying between 0% and 10.05% with a mean of 2.45%. To summarize, Models 2 and 3 differed from Model 1 in the use of moderated regression; Model 3 differed from Models 1 and 2 in the use of response bias indicators rather than instructional set. Each accounted for a smaller proportion of variance than the previous model.

Holden (2008) interpreted these results as evidence that moderated regression using response bias indicators tends to underestimate the occurrence of bias. The implication is that response bias can be a substantially greater problem than the research literature using this strategy would suggest. Before such a conclusion is accepted, however, it is important to place the results of this study in the context of prior discussions of moderated regression, and to note certain limitations of the study that could have limited the effectiveness of the bias indicators.

2. Moderated regression

Based on the smaller mean proportion of variance predicted by Model 2 versus Model 1, it was concluded that the moderated regression results underestimate the proportion of variance predicted by response bias. This conclusion is based on several assumptions. One is that Model 1 provides the "true" estimate of that quantity. In fact, it is a well-known problem in the use of standardized effect size measures that different statistics validly can lead to different conclusions about the strength of an effect (e.g., McGrath & Meyer, 2006). There is no authoritative rationale for awarding precedence to the results of Model 1 over those from Model 2. However, even if this point is acceded, the conclusion requires the further assumption that the proportion of variance values generated for Model 1 are comparably scaled with those generated by Models 2 and 3. Despite the reference to "proportion of variance" in both statistical models, the lower values for Model 2 than for Model 1 suggests that perhaps they are not. In fact, there is a literature that would support this conclusion as well.

McClelland and Judd (1993) attempted to explain why it is that moderator terms are difficult to detect in field studies (see also O'Connor, 2006). They drew a number of conclusions, the most relevant of which was that the proportion of variance predicted by an interaction term can look trivial and yet still be significant and even important (see also Evans, 1985). This occurs because the variance of the hierarchical term is reduced by properties of the joint distribution of the two raw predictors that are not considered in Holden's (2008) Model 1. Several studies have concluded that moderator terms in field studies rarely account for more than 3% of criterion variance (e.g., Chaplin, 1991). It is noteworthy that these discussions have usually focused on the squared partial correlation, which would actually tend to be slightly larger than the squared part correlations reported by Holden (2008).

To explore this issue further, power analysis can be used to offer a different perspective on the relative scaling of Models 1 and 2. For this purpose, the difference between independent correlations (Model 1) is usually represented by Cohen's (1988) q statistic, while the incremental validity of the moderator term (Models 2 and 3) can be evaluated using f^2. Using values from Holden's (2008) tables, q values were estimated directly. The squared part correlations were then divided by (1 – the squared part correlation). The result of this computation provides an estimate of f^2. This computation actually underestimates the true value of f^2, since the correct denominator should be (1 – the squared multiple correlation for the full model). However, the extent of the underestimation should generally be small unless the predictors prove to be unusually powerful. Since disparities in the size of the two samples generating the correlations is considered in the estimation of power for the z test but not the multiple regression F test, group sizes were equalized for the power analysis of q to eliminate differences due to skew in the grouping variable.

The weighted mean q value across eight comparisons was .392. For the sample sizes used in this study, power analysis software (Faul, Erdfelder, Lang, & Buchner, 2007) indicated this translates into a mean power of .80. The weighted mean f^2 estimate for Model 2 was .056, which represents a mean power of .93. In other words, based on the study of power analysis, the Model 2 estimates of effect size were relatively larger than those presented for Model 1.[1]

This presentation is not intended to suggest the results of the power analysis offer a more valid representation of the relative size of effects from Model 1 versus Model 2. The point is that the two models are based on very different approaches to understanding the proportion of variance predicted by the moderator term. The two models do not necessarily allow for straightforward comparisons of one producing larger effects than the other.

3. Response bias indicators

The results for the power analysis of Model 3 were more consistent with Holden's (2008) conclusions. The mean f^2 estimate was .026, which translated into a mean power of .65. Several issues must be considered when interpreting these results. The most obvious is the inevitable decline in

[1] Though questions have been raised about the misuse of post-hoc power analysis (e.g., Hoenig & Heisey, 2001), those questions are not relevant in the context of comparing the power of different statistical models. Also, it should be noted this analysis ignores other factors that can affect power such as violation of the moderated regression assumption of homogeneous error variances (Alexander & DeShon, 1994).

effect size that must occur when a questionnaire is compared with perfectly reliable and valid classification as was used in Model 2. Several other issues are more specific to the studies presented. As the author noted, there was a floor effect for the predictor in the first study when the means of the standard and faking good groups were compared. The study also used measures that have only been validated as predictors of faking good to detect both faking good and faking bad.

In addition, by averaging effect sizes across five bias indicators in the first study, Holden (2008) treated questionnaire as if it were a random variable. Prior research suggests not all indicators of response bias in common use are of equal validity, however (e.g., Barthlow, Graham, Ben-Porath, Tellegen, & McNulty, 2002). In the first study, the moderator terms based on the two scales from the Balanced Inventory of Desirable Responding (Paulhus, 1998) never predicted more than 1.51% of criterion variance. The other three multiplicative terms were associated with substantially larger values, at least in the case of faking bad where the predictor did not demonstrate a ceiling effect (7.13–10.05%). This finding suggests the BIDR scales may have been less valid than the other indicators in this setting, though the cause for this difference is uncertain. Unfortunately, the only bias indicator used in the second study was one of the two BIDR scales, though it was the more effective of the two used in the first study. The results do not provide strong support for the conclusion that response bias indicators as a group are inadequate to the task.

4. Conclusions

Holden (2008) has offered some intriguing hypotheses for explaining why it is that response biases are so difficult to detect in field settings. Unfortunately, the statistical analysis demonstrates several flaws that compromise its interpretation. The results are consistent with concluding that at least some popular response bias indicators may be less valid than others, a conclusion that is also consistent with some prior research. The data presented are insufficient to generalize this conclusion to bias indicators as a class, however. The case for attributing some of the problem to moderated regression is weaker, particularly when discussed in light of previous literature on effect sizes for product terms. In particular, this argument ignores prior literature finding that Model 1 may be no better than moderated regression at identifying response bias in field settings (e.g., Piedmont et al., 2000). While Holden's hypotheses are intriguing, the statistical conclusion validity of this study must be closely questioned. It is frustrating to note that despite 60 years of study, the importance of response bias in applied settings remains unresolved.

Acknowledgement

I am grateful to Stephen Armeli for pointing me to materials on evaluating moderator effects.

References

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin, 115*, 308–314.

Barthlow, D. L., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & McNulty, J. L. (2002). The appropriateness of the MMPI-2 K correction. *Assessment, 9*, 219–222.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59*, 143–178.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Costa, P. T., Jr., & McCrae, R. R. (1991). *NEO Five-Factor Inventory—Form S*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475–494.

Evans, M. G. (1985). A Monte Carlo study of the effects of correlated method variance in moderated multiple regression analysis. *Organizational Behavior and Human Decision Processes, 36*, 305–323.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Gouvier, W. D., Lees-Haley, P. R., & Hammer, J. H. (2003). The neuropsychological examination in the detection of malingering in the forensic arena: Costs and benefits. In G. P. Prigatano & N. H. Pliskin (Eds.), *Clinical neuropsychology and cost outcomes research: A beginning* (pp. 405–424). New York: Psychology Press.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19–24.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270–1285.

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences, 44*, 311–321.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*, 376–390.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods, 11*, 386–401.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*, 1094–1102.

O'Connor, B. P. (2006). Programs for problems created by continuous variable distributions in moderated multiple regression. *Organizational Research Methods, 9*, 554–567.

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245–269.

Paulhus, D. L. (1998). *Paulhus Deception Scales (PDS) user's manual*. North Tonawanda, NY: Multi-Health Systems.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582–593.

Rorer, L. G. (1965). The great response style myth. *Psychological Bulletin, 63*, 129–156.